Zpracování a kategorizace česky psaných textových dokumentů neuronovou sítí

Pavel Mautner, Roman Mouček

Abstrakt: *Kohonenova samoorganizující mapa byla navržena pro shlukování vstupních vektorů a mapování spojitého vícerozměrného signálu do diskrétního prostoru nižších dimenzí (nejčastěji 2D). Jednou z mnoha oblastí, ve kterých může být tato mapa využita, je i oblast zpracování textových dokumentů. V rámci projektu WEBSOM byla vytvořena řada metod založených na Kohonenově mapě. Tyto metody jsou vhodné jak pro vyhledávání informací v textových dokumentech, tak pro organizaci velké kolekce textových dokumentů. Metody byly testovány na kolekci anglicky a finsky psaných dokumentů. Tento článek se zabývá aplikací metody WEBSOM na kolekci česky psaných dokumentů. Je zde popsán základní princip metody, způsob převodu textové informace na číselnou reprezentaci zpracovávanou Kohonenovou mapou. Alternativně s Kohonenovou mapou byla testována i Carpenter-Grossbergova ART-2 síť, běžně používaná pro adaptivní shlukování vstupních vektorů. Výsledky dosažené s využitím této sítě jsou rovněž prezentovány v tomto článku.*

Abstract: *The Kohonen Self-organizing Feature Map (SOFM) has been developed for the clustering of input vectors and for projection of continuous high-dimensional signal to discrete low-dimensional space. The application area, where the map can be also used, is the processing of text documents. Within the project WEBSOM, the some methods, based on SOFM have been developed. These methods are suitable either for text documents information retrieval or for organization of large document collections. All method have been tested on collections of english and finish written documents. This article deals with application of WEBSOM methods for czech-written documnts collections. The basic principles of WEBSOM methods, transformation of text information into the real components feature vector and results of documents classification are described in the article. The Carpenter-Grossberg ART-2 neural network, normally used for adaptive vector clustering, was also tested as a document categorization tool. The results achieved by using of this method are also presented here.*

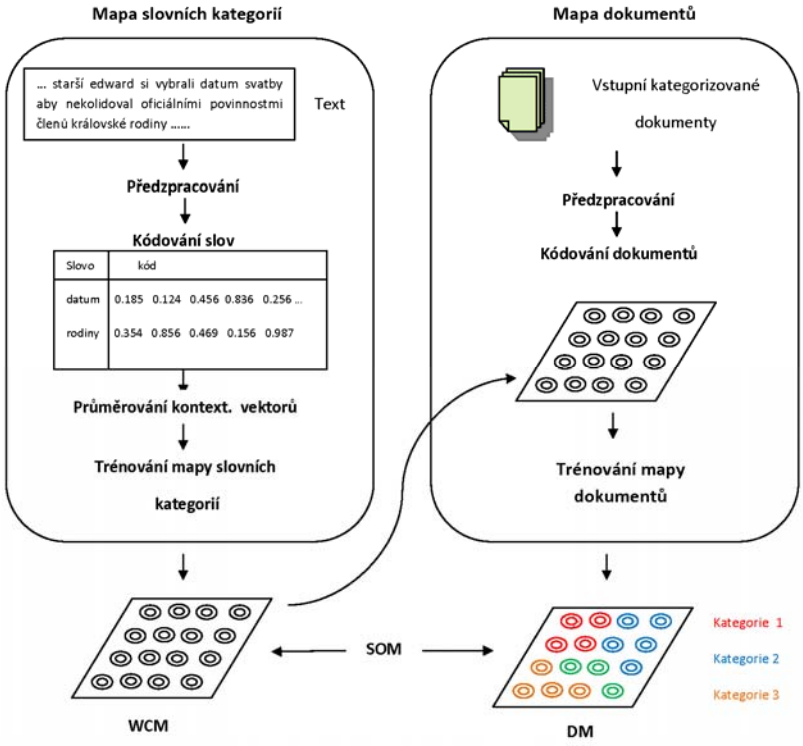
V dnešní době jsou dostupné stále větší kolekce dokumentů v elektronické podobě. Nalezení relevantních informací v takto obsáhlých kolekcích, dostupných převážně na internetu, je mnohdy obtížný a časově náročný proces. Mezi základní způsoby vyhledávání informací v dokumentech patří metoda založená na položení vhodného dotazu (např. dotazu obsahujícího klíčová slova z vyhledávané problematiky) a následném porovnávání obsahu dokumentu s klíčovými slovy obsaženými v dotazu. Vzhledem k tomu, že přirozený jazyk nám umožňuje určitou volnost v použití slov při kladení dotazu (např. použití synonym), může nastat situace, kdy seznam vrácených dokumentů obsahujících klíčová slova z položeného dotazu je velký a obsahuje celou řadu nerelevantních dokumentů (popř. odkazů na dokumenty).

Jedním ze způsobů urychlení vyhledávání informací v obsáhlých kolekcích je kategorizace dokumentů do několika tříd na základě tématu, o kterém daný dokument hovoří. Na základě slov obsažených v položeném dotazu je pak možné odhadnout třídu (doménu), které se daný dotaz týká a tím pádem zúžit vyhledávací prostor na dokumenty z dané doménové oblasti. Je zřejmé, že použitím tohoto mechanismu dojde jak ke snížení časové náročnosti vyhledávání dokumentů, tak k redukci seznamu odkazů na vyhledané dokumenty.

V minulosti byla navržena celá řada metod, které klasifikují dokumenty do daných doménových oblastí. Tyto metody však vyžadují vhodnou reprezentaci dokumentů uložených v databázi. Nejčastěji se k reprezentaci dokumentů používají klasické přístupy založené buďto na vektorovém modelu dokumentu, popř. na latentní sémantické indexaci [3]. Jedním z poněkud netradičních přístupů k reprezentaci dokumentů a jejich následné klasifikaci je metoda WEBSOM, založená na Kohonenově samoorganizující mapě [1]. Tato metoda byla navržena pro automatické zpracování a kategorizaci anglicky (popř. finsky) psaných dokumentů dostupných na internetu a následné vyhledávání informací v těchto dokumentech. Tento článek se zabývá využitím metody WEBSOM pro automatickou klasifikaci česky psaných dokumentů. V kapitole 2 je popsána základní architektura Kohonenovy samo-organizující sítě a architektura metody WEBSOM. Kapitola 3 se zabývá reprezentací dokumentů číselným příznakovým vektorem, způsobem vytvoření slovních kategorií a následnou klasifikací dokumentu. V kapitole 4 jsou uvedeny výsledky některých experimentů a další možné modifikace navržené metody.

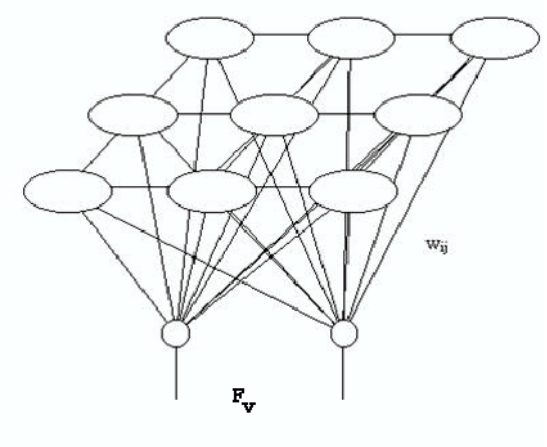
1. Architektura systému pro klasifikaci česky psaných dokumentů

Metoda WEBSOM je založená na dvouvrstvé architektuře znázorněné na obr. 1.



Obr. 1: Architektura WEBSOM

První vrstva zpracovává vstupní vektor reprezentující jednotlivá slova v dokumentu a vytváří tzv. mapu slovních kategorií WCM (**W**ord **C**ategory **M**ap), druhá vrstva, mapa dokumentů DM (**D**ocument **M**ap) provádí kategorizaci dokumentů na základě výstupu produkovaného mapou slovních kategorií. Obě zmíněné vrstvy WCM i DM jsou tvořeny Kohonenovou samoorganizující mapou (viz obr. 2), což je umělá neuronová síť, původně navržená ke shlukování vstupních dat a mapování spojitého vícerozměrného signálu do diskrétního prostoru nižších dimenzí (nejčastěji 2D).



Obr. 2: Architektura Kohonenovy mapy

Kohonenova mapa je složena z jedné vrstvy neuronů, obvykle uspořádaných do dvou-rozměrné mřížky. Každý neuron výstupní vrstvy je propojen přes váhový vektor w_{ij} s jednotlivými komponentami vstupního vektoru. Jednotlivé neurony pak počítají výstupní odezvu podle následujícího vztahu:

$$d_j(t) = \sum_{j=1}^{n-1} \left(x(t) - w_{ij}(t)\right)^2$$

kde t časový okamžik, ve kterém sledujeme výstup, $x_i(t)$ jsou komponenty vstupního vektoru a $w_i(t)$ je váhový vektor neuronu, jehož výstup sledujeme. Následně je vybrán neuron, jehož výstup *d* má nejmenší odchylku od vstupního vektoru. Tento neuron je označen jako vítězný neuron (BMU- Best Matching Unit). Výběr BMU je zajištěn pomocí laterálních spojů mezi neurony výstupní vrstvy. Vítězný neuron definuje odezvu sítě na vstupní vektor. Tato odezva je využita jak v procesu trénování, kdy jsou postupně nastavovány váhy vítězného neuronu a neuronů v jeho okolí, tak v procesu klasifikace, kdy vítězný neuron určuje třídu, do níž je zařazen vstupní vektor. Podrobný popis Kohonenovy mapy, včetně algoritmů trénování lze nalézt např. v [2].

2. Reprezentace dokumentů číselným vektorem a trénování mapy dokumentů

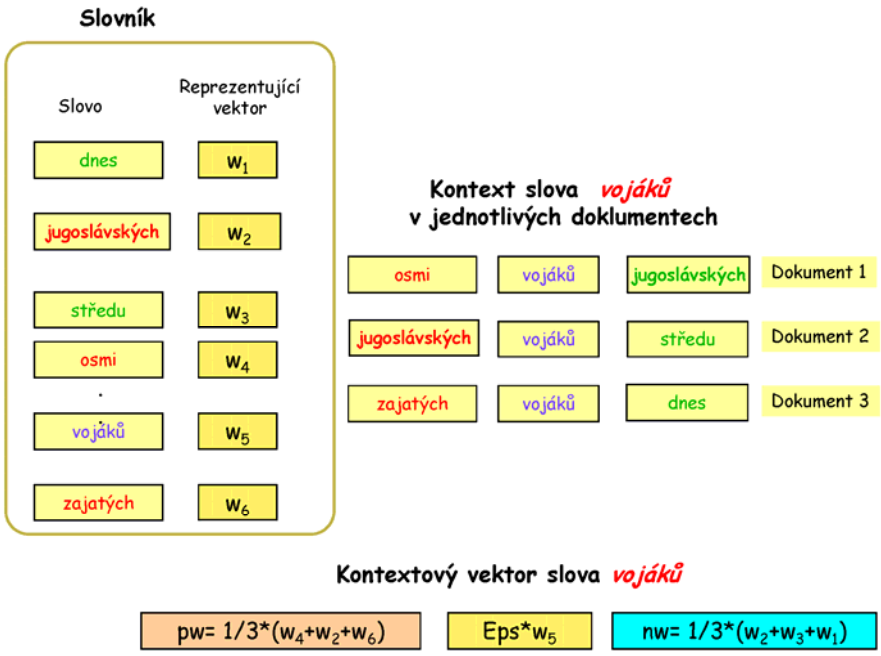
V předchozí kapitole byla naznačena architektura systému pro zpracování a kategorizaci dokumentů. Vzhledem k tomu, že vstupní vrstva systému je tvořena Kohonenovou mapou, která zpracovává číselné vektory, je nutné vhodným způsobem transformovat vstupní text na odpovídající číselný vektor. Jedním ze způsobů, jak kódovat dokument, je použití tzv. vektorového modelu [3]. Předpokládejme, že máme doménovou oblast obsahující celkem n slov. Jednotlivé dokumenty z této doménové oblasti mohou být kódovány n-rozměrným vektorem, jehož komponenty reprezentují jednotlivá slova z doménové oblasti. Přítomnost daného slova v dokumentu je ve vektorovém modelu vyjádřeno nastavením odpovídajícího prvku vektoru na jednotku (popř. na číselnou hodnotu reprezentující četnost výskytu slova v daném dokumentu). Je zřejmé, že tento způsob reprezentace dokumentů je nevhodný z důvodů paměťové náročnosti (rozměr vektoru je shodný s počtem slov v dané doménové oblasti) a s tím související i náročnosti časové.

V [1] byl popsán způsob reprezentace dokumentu vektorem slovních kategorií. Tento vektor má délku shodnou s počtem výstupních neuronů mapy WCM, a je vytvářen na základě kontextu, v jakém se jednotlivá slova vyskytují v dokumentech. Vektor slovních kategorií je vytvořen pro daný dokument tak, že se jednotlivá slova dokumentu postupně předkládají natrénované mapě WCM a sleduje se odezva mapy, tj. nalezne se BMU pro daný vstup a ve vektoru slovních kategorií se inkrementuje položka shodná s pozicí BMU v mapě WCM. Mapa WCM je trénována kontextovými vektory cw_i, které jsou vytvořeny následujícím způsobem:

1. Každému slovu ve slovníku pro danou doménovou oblast je přiřazen jednoznačný náhodný n-prvkový vektor w_i (tzv. reprezentující vektor), jehož prvky jsou reálná čísla.
2. Prohledají se zpracovávané dokumenty a naleznou se všechny výskyty zpracovávaného slova (reprezentovaného vektorem w_i).
3. Je nalezen kontext, ve kterém se slovo w_i nachází, tj. vezme se m-slov (v našem případě m=1), která předchází, popř. následují zpracovávané slovo w_i a z takto nalezených slov se určí hodnoty pw_i (průměrný vektor stanovený z reprezentujících vektorů všech slov dokumentu, která se vyskytují před slovem w_i) a nw_i (průměrný vektor stanovený z reprezentujících vektorů všech slov dokumentu, která se vyskytují za slovem w_i).
4. Kontextový vektor slova cw_i je vytvořen z hodnot pw_i, w_i, nw_i následovně:

$$cw_i = \begin{bmatrix} pw_i \\ \varepsilon w_i \\ nw_i \end{bmatrix},$$

kde ε je váha reprezentujícího vektoru slova w_i. Na obr. 3 je znázorněno vytvoření kontextového vektoru pro slovo „vojáků“.

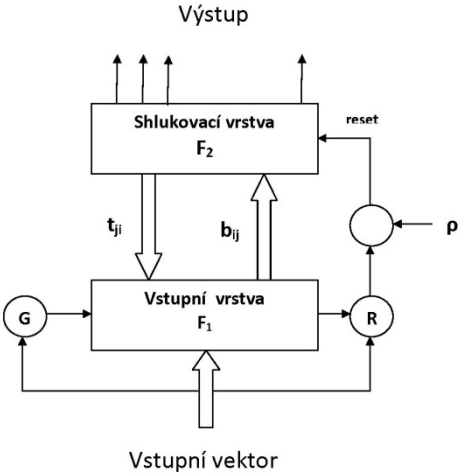


Obr.3: Vytváření kontextového vektoru pro slovo „vojáků“.

Je zřejmé, že slova, která se ve zpracovávaném dokumentu vyskytnou v podobném kontextu, budou mít i podobný kontextový vektor a dá se tedy říct, že budou patřit i do stejné slovní kategorie. Na základě tohoto předpokladu je možné natrénovat mapu dokumentů.
Mapa dokumentů DM (obr. 1) zpracovává výstup WCM (což je vektor slovních kategorií) a na základě tohoto výstupu provádí zařazení dokumentů do jednotlivých kategorií. Lze předpokládat, že dokumenty s podobným obsahem budou mít podobné vektory slovních kategorií. Na základě tohoto předpokladu je možné trénovat mapu DM pro kategorizaci dokumentů. Vzhledem k tomu, že Kohonenova mapa je trénovaná bez učitele, vytvoří se během trénování DM pouze shluky podobných dokumentů a jednotlivé významové kategorie (tématické obsahy dokumentů) je nutné přiřadit jednotlivým neuronům až po natrénování sítě.

3. Kategorizace dokumentů neuronovou sítí ART

V předchozí kapitole byl popsán způsob kategorizace dokumentů využívající Kohonenovu mapu. Jak již bylo řečeno, výstupem mapy DM jsou shluky obsahově podobných dokumentů a těmto shlukům je nutné po natrénování mapy přiřadit příslušné kategorie. To může být v některých případech značně komplikované, neboť se dá jen obtížně stanovit, kde se v mapě dokumentů nachází přesné hranice mezi shluky a tím i mezi jednotlivými kategoriemi.
Tento problém je možné řešit volbou jiného typu neuronové sítě, která má podobné vlastnosti jako Kohonenova mapa, ale jejímž výstupem jsou přesně stanovené kategorie. Pokud bychom přesně znali počet jednotlivých kategorií a měli bychom předem určené kategorie textových dokumentů, bylo by možné zvolit některou z umělých neuronových sítí trénovaných s učitelem. Vzhledem k tomu, že v mnoha případech nám jde o rozdělení dokumentů do jednotlivých kategorií pouze na základě podobnosti tématu, o kterém pojednávají, byla zvolena neuronová síť ART, která je opět založena na shlukování vstupních vektorů, jejím výstupem je však přímo informace o třídě, do které je vstupní vektor zařazen.



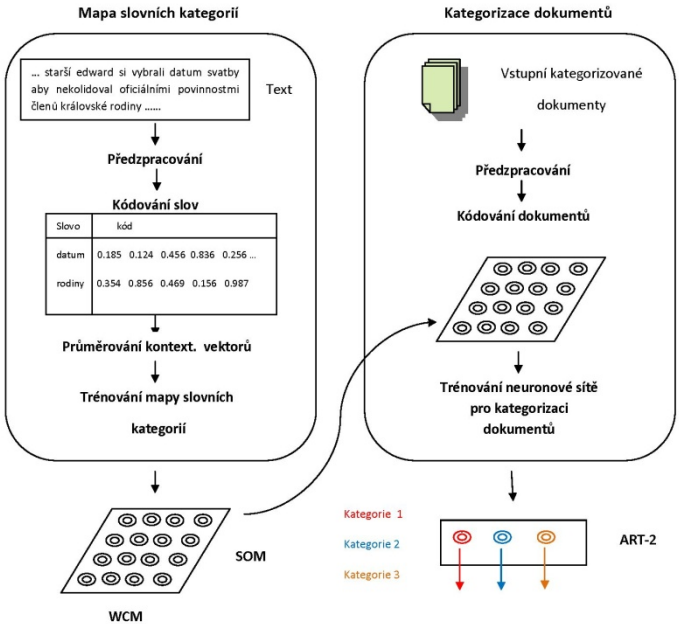
Obr. 4: Architektura sítě ART-2

Síť ART (Adaptive Resonance Theory) byla navržena Carpentrem a Grossbergem ke shlukování vstupních vektorů. Existuje několik typů sítě ART (ART-1, ART-2, ARTMAP) vzájemně se lišících architekturou a typem vstupních vektorů, které jsou schopné zpracovat. V našem případě byla použita síť ART-2, která shlukuje vstupní vektory s reálnými složkami. Zjednodušená architektura sítě je znázorněna na obr. 4. Síť se skládá ze dvou vrstev neuronů

označených $\mathbf{F_1}$ (tzv. vstupní vrstva) a $\mathbf{F_2}$ (tzv. shlukovací vrstva), které jsou vzájemně propojeny a pomocných neuronů **G** a **R**, které slouží k řízení činnosti sítě a vytváření jednotlivých shluků. Počet neuronů ve vstupní vrstvě $\mathbf{F_1}$ je roven rozměru vstupního vektoru, počet neuronů ve shlukovací vrstvě $\mathbf{F_2}$ je roven maximálnímu povolenému počtu shluků. Propojení vrstev $\mathbf{F_1}$ a $\mathbf{F_2}$ je realizováno přes váhové vektory $\mathbf{b_{ij}}$ a $\mathbf{t_{ji}}$, ve kterých jsou uloženy vzorové obrazy jednotlivých shluků, které se mohou adaptivně měnit v závislosti na příchozím vstupním vektoru. Vlastní popis činnosti sítě a algoritmus jejího trénování je poněkud komplikovaný a přesahuje rámec tohoto článku. Podrobnější informace jsou uvedeny např. v [4]. Stručně lze činnost sítě shrnout do následujících bodů:

1. Vstupní vektor sítě je porovnán se vzorovými vektory (uloženými ve vahách spojení, vedou z vrstvy $\mathbf{F_1}$ do $\mathbf{F_2}$).
2. Neuron vrstvy $\mathbf{F_2}$ s největší odezvou je zvolen jako vítězný neuron je ověřeno, zda podobnost mezi vstupem a váhovým vektorem splňuje předem nastavené kritérium, které lze modifikovat změnou parametrem ρ.
3. Pokud je podobnost dostatečná zařadí se vstupní vektor do odpovídajícího shluku a adaptují se váhy neuronu vrstvy $\mathbf{F_2}$, který tento shluk reprezentuje
4. Pokud není nastavené kritérium splněno, zablokuje se vítězný neuron a činnost se opakuje od bodu 2, dokud není nalezen neuron splňující dané kritérium.
5. Pokud jsou zablokovány všechny neurony (tj. žádný shluk nesplňuje kritérium podobnosti mezi vzorovým prvkem shluku a vstupním vektorem), je vytvořen nový shluk a vstupní vektor se stává jeho vzorovým obrazem (dochází k adaptaci vah nově aktivovaného neuronu).

Modifikovaná architektura systému zpracování dokumentů, který ke kategorizaci využívá síť ART-2, je znázorněna na obr. 5.

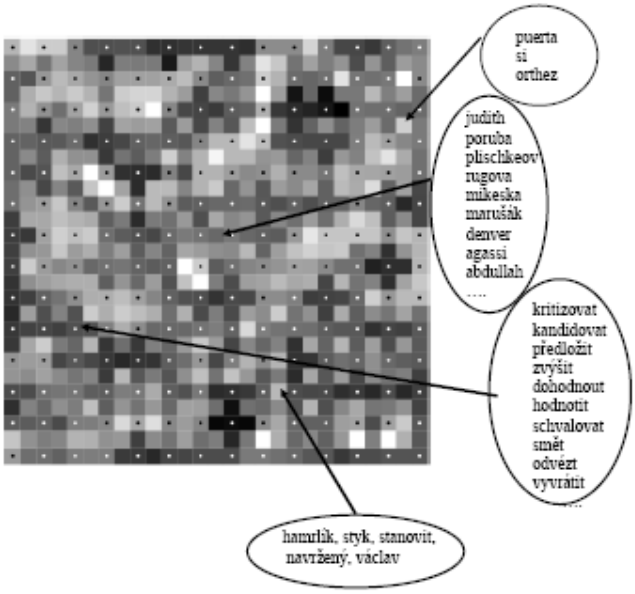


Obr. 5: Kategorizace dokumentů sítí ART-2

Jednotlivá slova dokumentu jsou opět nejprve zpracována natrénovanou mapou WCM a výsledný vektor slovních kategorií dokumentu je přiveden na vstup sítě ART-2, která zařadí vstupní vektor do odpovídajícího shluku, popř. vytvoří shluk nový, odpovídající nové kategorii vstupních dokumentů.

4. Dosažené výsledky a závěr

Systém pro kategorizaci dokumentů popsaný v předchozích kapitolách, byl testován na korpusu dokumentů obsahujících tiskové zprávy ČTK. Celkem bylo k dispozici 7600 dokumentů z 6 tématických okruhů (kategorií), obsahujících celkem 145 766 slov (nevýznamová slova, byla z dokumentů odstraněna). Simulace Kohonenovy mapy byla provedena jednak dostupnými simulátory (SOM-PAK, SOM-toolbox), jednak vlastní implementací. Obě Kohonenovy mapy byly trénovány nejprve sekvenčním algoritmem trénování, ale vzhledem k jeho značné časové náročnosti trénování, byl nakonec implementován batch algoritmus trénování [], který dává podobné výsledky, je však mnohonásobně rychlejší. Testované dokumenty byly ručně kategorizovány do šesti tříd, podle tématu o kterém pojednávají (např. sport, politika, zákonodárství apod.).
Velikost mapy slovních kategorií (první vrstva systému) byla zvolena tak, aby v každé kategorii bylo průměrně 25 slov. Ze slovníku, který byl použit k trénování, byla před vygenerováním reprezentujících vektorů odstraněna slova, jejichž četnost výskytu v dokumentu byla nižší než předem zvolený práh.
Mapa kategorií dokumentů (druhá vrstva systému z obr. 1) byla vytvořena z devíti neuronů uspořádaných do dvourozměrné mřížky 3x3. Vstup mapy kategorií dokumentů tvořil výstup mapy slovních kategorií, který byl předzpracován Gausovou konvoluční maskou.
Na obrázku 3 je zobrazena mapa slovních kategorií trénovaná jednotlivými slovy z množiny 100 dokumentů. Z obrázku je patrné zřejmé, že některé výstupní jednotky mapy reagují na slova, která odpovídají určitým syntaktickým kategoriím (např. slovesa, popř. vlastní jména), jiné jednotky vytváří odezvu na slova z různých syntaktických kategorií. Tento jev je dán vlastnostmi daného přirozeného jazyka a způsob, jak ho odstranit, bude předmětem dalšího zkoumání.



Obr. 3: Mapa slovních kategorií

Číslo neuronu v mapě dokumentů DM (kategorie)	Podíl dokumentů v % přiřazených danému neuronu s tématickým okruhem			
	sport	politika	zákonodárství	společnost
1	1	7	1	0
2	18	4	2	1
3	8	1	0	0
4	9	11	3	0
5	0	0	0	0
6	4	11	5	3
7	0	0	0	0
8	0	0	0	0
9	0	9	1	1

Tabulka 1: Výsledek kategorizace dokumentů metodou WEBSOM

V tabulce 1 jsou uvedeny výsledky klasifikace dokumentů pro vytvořenou mapu slovních kategorií z obr. 3. Testované dokumenty byly v tomto případě pouze ze 4 tříd a obsahovaly následující témata: sport, politika, zákonodárství a společnost. Přiřazení odpovídající kategorie jednotlivým neuronům, bylo provedeno až po natrénování sítě.
V další fázi byly provedeny testy se systémem, ve kterém je jako kategorizátor použita síť ART-2 (viz obr. 5).

Číslo shluku v ART-2 (kategorie)	Podíl dokumentů v % přiřazených danému shluku s tématickým okruhem			
	sport	politika	zákonodárství	společnost
0	2.6	55.6	14.9	26.7
1	0.5	2.7	0.9	2.2
2	0.8	5.8	21.3	15.6
3	19.9	5	7.2	20
4	8.8	0.5	0.5	0
5	2.2	5.0	17.2	2.2
6	33	3.9	14	8.9
7	1	0	0	0
8	10.1	0	2.3	0
9	8.0	18.9	0.9	8.9
N	13.1	2.5	20.8	15.6

Tabulka 2: Výsledek kategorizace dokumentů sítí ART-2

Na první pohled je zřejmé, že výsledky kategorizace dokumentů nejsou příliš přesvědčivé a jsou bohužel do značné míry ovlivněny výstupem mapy slovních kategorií. Je tedy nutné se v první řadě zaměřit na modifikaci této mapy s cílem dosáhnout co nejlepších výsledků při kategorizaci slov. V další fázi bude také ověřena možnost náhrady mapy kategorií dokumentů jinou neuronovou sítí (ART-2, popř. některou jinou sítí, využívající učení s učitelem vícevrstvý perceptron, LVQ, apod.).

Poděkování:

Tato práce vznikla v rámci řešení projektu MŠMT č. 2C06009 „Prostředky tvorby komplexní báze znalostí pro komunikaci se sémantickým webem v přirozeném jazyce“.

Literatura:

[1] Kaski, S., Honkela, T., Lagus, K., and Kohonen, T.: *WEBSOM-self-organizing map of document collections*, Neurocomputing 21 (1998) 101-117
[2] Kohonen, T.: *Self-Organizing Map*, Springer-Verlag, Berlin Heidelberg, 2001
[3] Manning, C.D., Raghavan, P., Schűtze, H.: *Introduction to Information Retrieval,* Cambridge university Press, 2007

Adresa:
Ing. Pavel Mautner, Ph.D.
Ing. Roman Mouček, Ph.D.
Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni, Univerzitní 8, 306 14 Plzeň
Email: mautner@kiv.zcu.cz, moucek@kiv.zcu.cz